**Artificial Intelligence Linguistic Developments and Written Assessments**

**Sharina Menke**

**California State University, Fullerton**

**Introduction**

When most people hear the word artificial intelligence, they immediately think of robots, like in the movie Bicentennial Man or I, Robot. Artificial intelligence, according to the Oxford dictionary, is "the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages." Applications of AI technology include: auto-journalism, online dating, smart phones, stocks and shares dealing, passport control, legal and financial services, and of course, education (Holmes, 2020).

There are 3 sub-fields of AI: Rule based systems, statistical algorithms, and machine learning. Computer science is particularly making vast and rapid advancements in the field of machine learning. Within the subfield of machine learning, *supervised* learning is when the AI can label new data, and *unsupervised* learning is when AI finds patterns in unlabeled data so that the AI can classify new data. Neural networks are based on animal brains, but are not nearly as complex. These networks usually involve only a few thousand electronic “neurons,” while the human brain has around 100 billion neurons. Deep learning involves neural networks that have many hidden layers. An example of this is the software Facebook uses to recognize faces. Types of machine learning include reinforcement learning (the system is continuously learning with success being rewarded) and generative adversarial networks (two deep neural networks competing against each other) (Holmes, 2020).

So exactly how *intelligent* is artificial intelligence? AI is not better than human beings at many tasks, and it is easy to fool. For example, AI is good at processes that can be challenging to humans, such as handling complex computations, but it is weak at processes that humans find easy, such as recognizing faces on the first try. This is known as Moravec's Paradox. Many researchers have proposed that the capabilities of machine learning may hit a ceiling. Virtual agents are also unlikely to get significantly better even with more data. Researchers are looking at merging rule-based AI and machine learning to enhance its capabilities. AI also needs humans to input the data it needs to do its job (A human labels the data, designs the network, trains the network, curates the output and makes value judgments) (Holmes, 2020).

Artificial Intelligence has been used in education for over 50 years. AIED is a huge industry with heavily funded companies such as SquirrelAI and Carnegie Leaf. Amazon, Google, and Facebook are investing big bucks in the field as well and the industry is projected to be worth 6 billion dollars in 2024.

AI technologies are becoming more common in classrooms. Student facing AI involves intelligent tutoring systems (ITS) which are based on the students response, dialogue based tutoring, exploratory learning environments such as Betty's Brain and Crystal Island, and automatic writing evaluations such as Open Essayist, E-rater, and PEG, which will be discussed later in this paper.

There are many AI-driven writing evaluation software that are being used to make writing suggestions, revise, or assess written assignments. The ethics of using machine readers for scoring have been challenged and research has revealed that these technologies are not nearly as effective as a human reader when it comes to assessing writing skills.

**An Assessment of Current Writing Evaluation Software**

Some writing evaluation software include: Open Essayist, E-rater, and PEG. Writing evaluation software can either be classified as formative (providing suggestions) or summative, which is an automatic revision of your written report.

OpenEssayist is a formative, automated, interactive feedback system that provides support for students as they write essays for summative assessment, which are writing critiques that are administered at the end of a learning assessment. The OpenEssayist system includes: “1) a linguistic analysis engine and 2) a web application that generates feedback for students” (Whitelock, et al., 2015). This software works with graph-based frequency to automatically extract key words, phrases and sentences from student essays (Whitelock, et al., 2015). A paper entitled “OpenEssayist: An Automated Feedback System that Supports University Students as They Write Summative Essays” delved into the effectiveness of the visual representation output by the OpenEssayist system. Effectiveness of these visual representations were measured by how well a group of users understood them and if they had meaning to the users. The study reported that: “A mixture of texts and graphics were understood by the participants with respect to the structure of the essay, i.e., the introduction, discussion and conclusion” (Whitelock, et al., 2015). According to their findings, the graphs and visual organizers facilitated critical reflection on the manner by which key words and phrases were being used. The AI system and visuals also detected whether the draft essays were adequately addressing the assignment question (Whitelock, et al., 2015). However, the more complex visualisations were more difficult for these naive users to comprehend. Suggestions from those involved in the study included adding more interactive elements so that the visualisations could be manipulated and thus more deeply comprehended by users. Visual metaphors with a layer of interactivity that allows for deeper analysis can effectively help students improve on these writing assessments.

The e-rater engine is an ETS (Educational Testing Service) capability that identifies features related to aspects of writing proficiency in student essays so they can be used for scoring and feedback. The e-rater engine is also used in the Criterion® Online Writing Evaluation Service. In a research report entitled: “Evaluation of e‐rater® for the Praxis I® Writing Test,” AI-powered scoring models were trained and evaluated for the essay task in the Praxis I® writing test. The Praxis I assessment is a teacher licensure test that is comprised of mainly multiple‐choice (MC) sections, but there is a writing portion of the English comprehension section (Ramineni et al, 2015). In this study, evaluation statistics, which included quadratic weighted kappa, Pearson correlation, and standardized differences in mean scores, were examined to determine how the e-rater model performed as a writing assessment tool compared to human readers (Ramineni et al, 2015). This study resulted in a generic e‐rater scoring model being recommended for use as an assessment tool in the Praxis I program due to the agreement threshold being within 1.5 units of the first human score (Ramineni et al, 2015).

**Ethical Issues of Using Artificial Intelligence in Written Assessments**

A group called Professionals Against Machine Scoring of Student Essays in High Stakes Assessment was established to combat what they thought was the wrongful use of AI technology to grade written student assignments in standardized testing. The group asserts that computers cannot effectively measure the essentials of written communication (Humanreaders.org, 2013).

According to their research, there are many ethical issues to consider when dealing with artificial intelligence being used to assess writing assignments. When students’ futures and livelihood depend on the scores they receive from standardized testing, serious consideration should be given to the shortcomings of using AI for assessing the written portion of the exams. Machine readers fall short in terms of measuring knowledge and skills displayed in writing (Perelman, 2012). Instead of being able to critique the complexities of the writing in an analytical manner, machines use absurdly reductive algorithms. When it comes to assessing sophistication of vocabulary, machine readers focus on the average length or frequency with which words are used, while development of ideas is reduced to the number of sentences per paragraph (Perelman, 2012).

The unifying theme within the essays present in A Sense of Audience in Written Communication is that “social context in which the writing takes place or the dynamic process inherent in the writer-reader relationship” cannot be ignored and it plays a crucial role in the quality of the writing (Rubin, et. al.,1990). The awareness that the audience is not a human being and rather, a machine that does not take into account the nuances of language, will surely alter the quality of the writing.

University of Copenhagen researchers have discovered a way, using artificial intelligence to determine whether a ghostwriter has written a students written assignment. Studies have shown that the use of ghostwriters (cheating on written assignments by paying a ghostwriter to write the assignment), have become widespread. High schools in Denmark currently use the Lectio platform to check if students have plagiarized from a previously submitted assignment. The issue with this platform is that Lectio does not detect if a student has hired a ghostwriter. The program, Ghostwriter, uses machine learning technology and Siamese neural networks to determine if a student has not written the report they submitted. According to this article, “the network is trained on large amounts of data to learn from representations of writing styles, which are then compared.” The ethics issue here is that the output given by Ghostwriter should not be used exclusively to convict a student of cheating, but only be used to substantiate a claim along with other evidence (University of Copenhagen, 2019).

**Conclusion**

Overall, human readers are more effective assessors of written assignments and tests than machine readers that are powered by artificial intelligence. Machine readers cannot accurately measure the aspects of intelligence that are displayed in written assessments: critical thinking, development of ideas, and complexity of vocabulary, among others. As such, human readers should assess written assignments for these qualities whenever possible, there should be transparency about when AI is being used as an assessment tool, and AI should be limited to evaluating grammar and punctuation and making writing suggestions.

**References**

Holmes, W. (2020). Artificial Intelligence in Education: Promise and Implications for Teaching and Learning [video lecture]. Retrieved from <https://vimeo.com/408368440/7836b675c5>

Perelman, L. (2012). Mass-market writing assessments as bullshit. In N. Elliot & L. Perelman (Eds.), Writing assessment in the 21st century: Essays in honor of Edward M. White (pp. 425-438). New York, NY: Hampton Press.

“Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment.” Humanreaders.org, 2013, humanreaders.org/petition/.

Ramineni, Chaitanya, Catherine S. Trapani, and David M. Williamson. "Evaluation of E‐rater® for the Praxis I® Writing Test." ETS Research Report Series 2015.1 (2015): 1-28. Web.

Rubin, D. L., & O'Looney, J. (1990). Facilitation of audience awareness: Revision processes of basic writers. In G. Kirsch & D. H. Roen (Eds.), A sense of audience in written communication (pp. 280-292). Newbury Park, CA: Sage.

Whitelock, Denise, Alison Twiner, John Richardson, Debora Field, and Stephen Pulman. "OpenEssayist: A Supply and Demand Learning Analytics Tool for Drafting Academic Essays." Proceedings of the Fifth International Conference on Learning Analytics and Knowledge 16-20 (2015): 208-12. Web.

University of Copenhagen. (2019, May 29). Tempted to cheat on a written exam? Artificial intelligence is 90% certain to nab you. ScienceDaily. Retrieved May 19, 2020 from www.sciencedaily.com/releases/2019/05/190529145048.htm